Report Number: WUCS-97-39

1997-01-01

Computational Detection of CpG Islands in DNA

Eric C. Rouchka, Richard Mazzarella, and David J. States

Regions of DNA rich in CpG dinucleotides, also known as CpG islands, are often located upstream of the transcription start side in both tissue specific and housekeeping genes. Overall, CPG dinucleotides are observed at a density of 25% the expected level from base composition alone, partially due to 5-methylcytosine decay (Bird, 1993). Since CpG dinucleotides typically occur with low frequency, CpG islands can be distinguished statistically in the genome. Our method of detecting CpG islands involves a heuristic algorithm employing classic changepoint methods and log-likelihood statistics. A Java applet has been created to allow for user interaction and visualization... **Read complete abstract on page 2.**

Computational Detection of CpG Islands in DNA

Eric C. Rouchka, Richard Mazzarella, and David J. States

Complete Abstract:

Regions of DNA rich in CpG dinucleotides, also known as CpG islands, are often located upstream of the transcription start side in both tissue specific and housekeeping genes. Overall, CPG dinucleotides are observed at a density of 25% the expected level from base composition alone, partially due to 5-methylcytosine decay (Bird, 1993). Since CpG dinucleotides typically occur with low frequency, CpG islands can be distinguished statistically in the genome. Our method of detecting CpG islands involves a heuristic algorithm employing classic changepoint methods and log-likelihood statistics. A Java applet has been created to allow for user interaction and visualization of the segmentation resulting from the changepoint analysis. The model is tested using several sequences obtainable from GenBank (NCBI, 1997), including a 220 Kb fragment of human X chromosome from the filanin (FLM) gene to the glucose-6-phosphate dehydrogenase (G6PD) gene which has been experimentally studied (Rivella, et. al., 1995; E.Y. Chen, et. all., 1996). Preliminary results suggest a breakpoint segmentation that is consistent with observable manual analysis. About 56% of human genes have associated CpG rich islands (Antequera and Bird, 1993). By identifying the CpG islands, it is thought that regions of DNA coding for housekeeping or tissue-specific genes can be located (Antequera and Bird, 1993) even in the absence of transcriptional activity. Biological experiments searching for such genes can then be narrowed given the locations of the CpG islands.

COMPUTATIONAL DETECTION OF CpG ISLANDS IN DNA

Eric C. Rouchka, Richard Mazzarella, and
David J. States

WUCS-97-39

September 1997

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
Saint Louis, MO 63110

ecr@ibc.wustl.edu
rich@borcim.wustl.edu
states@ibc.wustl.edu

Abstract

Regions of DNA rich in CpG dinucleotides, also known as CpG islands, are often located upstream of the transcription start site in both tissue specific and housekeeping genes. Overall, CpG dinucleotides are observed at a density of 25% the expected level from base composition alone, partially due to 5-methylcytosine decay (Bird, 1993). Since CpG dinucleotides typically occur with low frequency, CpG islands can be distinguished statistically in the genome.

Our method of detecting CpG islands involves a heuristic algorithm employing classic changepoint methods and log-likelihood statistics. A Java applet has been created to allow for user interaction and visualization of the segmentation resulting from the changepoint analysis. The model is tested using several sequences obtainable from GenBank (NCBI, 1997), including a 220 Kb fragment of human X chromosome from the filanin (FLN) gene to the glucose-6-phosphate dehydrogenase (G6PD) gene which has been experimentally studied (Rivella, et. al., 1995; E.Y. Chen, et. al., 1996). Preliminary results suggest a breakpoint segmentation that is consistent with observable manual analysis.

About 56% of human genes have associated CpG rich islands (Antequera and Bird, 1993). By identifying the CpG islands, it is thought that regions of DNA coding for housekeeping or tissue-specific genes can be located (Antequera and Bird, 1993) even in the absence of transcriptional activity. Biological experiments searching for such genes can then be narrowed given the locations of the CpG islands.

COMPUTATIONAL DETECTION OF CpG ISLANDS IN DNA

Eric C. Rouchka, Richard Mazzarella, and David J. States
Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
Saint Louis, MO 63110
Ecr@ibc.wustl.edu
rich@borcim.wustl.edu
states@ibc.wustl.edu

September 1997

Introduction

Deoxyribonucleic acid, also known as DNA, is the genetic blueprint for life. DNA is composed of a linear chain of four nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). Information is encoded in the genome in independently heritable units known as genes. A gene typically includes control signals that determine when it will be active, a promoter that signals where the sequence should be copied into DNA, and a protein-coding region. There are two basic types of genes: housekeeping and tissue specific. Housekeeping genes are genes that are transcriptionally active (i.e. produce proteins) in cells throughout the body. Tissue specific genes, on the other hand, are transcriptionally active only in certain cells. Experimental results suggest that all housekeeping genes and 40% of the tissue specific genes in humans have an associated CpG island (Bird, 1993). It is proposed that by locating CpG islands in sequences of vertebrate DNA gene positions can be postulated. This paper will present characteristics of CpG islands in vertebrates and how they can be distinguished in a statistical fashion.

CpG Island Characteristics

Chemically, DNA is composed of nucleoside monomers ("bases") linked by a phosphate from the 3' hydroxyl of one sugar to the 5' hydroxyl of the next. CpG islands are regions of DNA high in the dinucleotide composition CG; that is, where a cytosine residue (C) is immediately followed by a guanine residue (G). The existence of CpG islands in vertebrates, particularly humans and mice, has been studied (Antequera and Bird, 1993; Aissani and Bernardi, 1991; Cross and Bird, 1995; Gardiner, 1996; Macleod, et. al., 1994). Aissani and Bernardi (1991) and Bernardi (1993) have studied the location of genes in the DNA of vertebrates and have grouped regions of chromosomes into isochores based on the nucleotide composition. It has been determined that both the majority of genes (Antequera and Bird, 1993; Gardiner, 1996) and CpG islands (Bernardi, 1993; Cross and Bird, 1995) are found on the Giemsa light or reverse bands of chromosomes, which are rich in the nucleotides C and G.

The CpG islands studied so far are mainly located upstream (5') of the gene that they are associated with, even though a few are located downstream (3') (Cross and Bird, 1995). Chen et. al. (1996) discuss this association by examining candidate genes occurring within a region of high G + C DNA. It is possible that CpG islands can be found in a region where no genes have previously been mapped. This information could help in setting up experiments to determine gene location.

CpG islands occur in unmethylated regions of DNA. Methylation is a process that adds a methyl group to a cytosine base. It offers a method to protect the DNA from restriction enzymes, as well as aiding in transcription regulation (BioTech Resources, 1996). It is interesting that CpG islands in vertebrates remain unmethylated. Three basic explanations suggest why CpG islands remain unmethylated (Macleod, et. al., 1994; Cross and Bird, 1995):

- DNA methyltransferase methylates GC-rich DNA poorly.
- CpG islands are methylated de novo, but the methylation is removed by an island-specific demethylating activity.
- Factors (such as Sp1 in mice) that deny access to DNA methyltransferase are bound to CpG islands.

Experimental evidence tends to support the latter two arguments.

Why CpG Islands can be Statistically Determined

If successive nucleotides in a DNA sequence were independent and identically distributed and residues occurred with equal frequency, it would be expected that by chance a nucleotide G or C would be observed at any given location 50% of the time. However, in genomic DNA, G + C occurs only 40% of the time. One simple method to find interesting regions of DNA would be to look for regions where the observed number of G's and the observed number of C's together exceeds 40%.

Since there are 4 different choices of nucleotides, it is expected that CpG dinucleotides will occur once in every 16 positions or 6.25% of the time by chance alone. As a result of the methylation process discussed in the previous section, CpG occurs at 25% the expected frequency (Bird, 1993). Over evolutionary time, this 5' methylcytosine decay has mutated the dinucleotide CpG into TpG (CpA on the complementary strand) so that both TpG and CpA are both over represented (Bird, 1980). A technique that Antequera and Bird (1993) use to locate possible CpG islands is to look at regions of DNA, at least 200 nucleotides in length, where the G + C content is at least 50% and an observed:expected CpG ratio is above 0.6. This criterion has also been used with the software package *CpG Isle* (Larsen, 1992; Lopez, NA) which characterizes CpG islands from sequences in the EMBL database. (*CpG Isle* can be obtained from the Internet at the URL ftp://ftp.ebi.ac.uk/pub/databases/cpgisle.)

CpG islands are also known as HTF islands (*HpaII* tiny fragments) since they are cut by the restriction enzyme *HpaII* (Cross and Bird, 1995). Other methods to experimentally determine the location of these islands include looking for rare-cutter sites and G/C boxes within DNA (Aissani and Bernardi, 1991). While these locations can be found experimentally in a wet lab, they can also be located using string-matching algorithms due to their specificity.

Segmentation Algorithm

As previously described, determining CpG island location by using the criterion that the G + C content is at least 50% and an observed:expected CpG ratio is above 0.6 will provide some clues as to where CpG islands will occur. However, such an approach can leave undetected CpG islands. It is also very specific to human nucleotide composition. A more sequence and organism independent approach is proposed that will help to detect even subtle CpG islands. Our aim is to implement this approach to search for other regions of compositional bias.

The problem can be approached as a classic changepoint problem (Carlin, et. al., 1992). Lawrence and Reilly (1985) have proposed changepoint methods to determine subsequence conservation within amino acid sequences using maximum likelihood estimation. Similar techniques can be used to determine the location of the breakpoints according to dinucleotide composition. The idea is to segment the DNA sequence into regions adjacent to one another with different CpG distribution.

First Phase: Breakpoint Segmentation

A heuristic approach has been taken to the changepoint problem in determining breakpoint locations. The general idea is to iterate a number of times, randomly choosing whether a new breakpoint should be tested, an existing one should be moved, or two adjacent regions should be merged. Each segment is assigned a score according to the formula in equation 1.

$$S = \overline{CpG} * \ln\frac{\overline{CpG}}{N} + CpG * \ln\frac{CpG}{N}$$

Equation 1: Segment Log-Probability Score

Here, S is the log probability score, $\overline{CpG}$ is the number of dinucleotides in the segment that are not CpG, CpG is the number of CpG dinucleotides in the segment, and N is the total number of dinucleotides in the segment. Note that $N = L - 1$ where L is the length of the segment in nucleotides. Table 1 indicates the dinucleotide counts for an example segment.

DNA is typically found in a double stranded conformation where one strand is complementary to the other and running in the opposite direction. Since it may not be known which strand the gene is transcribed from, both strands should be searched for dinucleotides. A nice property of the CpG dinucleotide is that its complement is the dinucleotide GpC. For the sequence ACGGTACGCGCGA, its complement is TGCCATGCGCGCT. The location of CpG islands in the complement should be looked for in the reverse direction due to the orientation. The CpG islands are as follows:

```
5′ ---- ACGGTACGCGCGA ---- 3′
3′ ---- TGCCATGCGCGCT ---- 5′
```

Note that the locations of CpG islands in both strands are identical. Thus, it is only necessary to search one strand for the location of CpG islands.

In order to determine whether or not a given breakpoint is significant, consider the diagram in figure 1.

Dinucleotide	Counts
AA	0
AC	2
AG	0
AT	0
CA	0
CC	0
CG	4
CT	0
GA	1
GC	2
GG	1
GT	1
TA	1
TC	0
TG	0
TT	0

Table 1: Dinucleotide Counts for Sequence ACGGTACGCGCGA

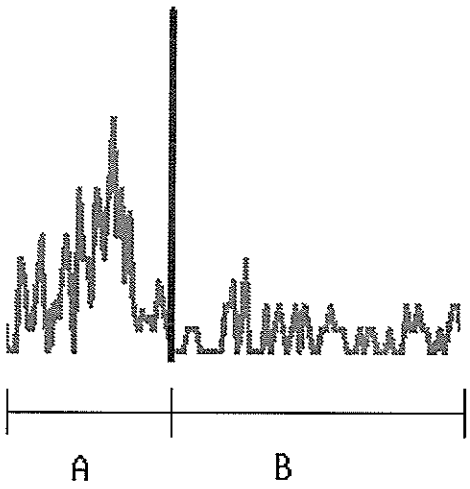

Figure 1: Breakpoint Segment Example

$$A_{score} = S \text{ for segment } A$$
$$B_{score} = S \text{ for segment } B$$
$$C_{score} = S \text{ for segment } A{+}B$$

For figure 1, S can be calculated using equation 1. If $A_{score} + B_{score} > C_{score} + Threshold$, then it is significant and a new breakpoint should be inserted at this location.

The threshold needs to be chosen in such a way as to ensure that all possible breakpoints are found, yet that no false breakpoints will result. It has been empirically determined that threshold values between 15 and 20 work best. It is also possible that the segmentation can over segment a CpG island. To overcome this problem, a post-processing step is invoked.

Second Phase: Post Processing

The purpose of the post-processing step is to further refine the boundaries of the segments found in the breakpoint segmentation phase. This can be accomplished in one of two ways. The first method is to merge segments together using a lower threshold value. The second method is to determine if two adjacent segments should be merged by determining if they are both above or both below the expected dinucleotide content based on the composition of the DNA sequence being studied. This in effect reverts back to the previous method of testing an observed:expected CpG ratio. Since this is done as a post-processing step, subtle islands will not be missed. By processing the breakpoints in this manner, false positives and fractionation of segments can be eliminated without loss of the true positives.

Generalization of the CpG Detection Algorithm

Location of CpG islands is only one application of the segmentation algorithm. Equation 1 can be easily changed to allow the user to determine breakpoints in other biologically significant locations. The user is given the option of finding breakpoints according to the C + G content (for the purpose of isolating isochores), mononucleotide content, purine/pyrimidine content (for structural purposes), and dinucleotide content. Equation 2 shows a generalization of equation 1.

$$S = \sum_{i=1}^{K} C_i * \ln F_i$$

Equation 2: Generalization of Segment Log-Probability Score

Here, K is the number of different compositions to segment by, C_i is the count of items in the segment of composition i and F_i is the frequency of items in the segment of composition i.

Java Applet Interface

A Java applet interface has been developed using Sun's JDK 1.1.1. It can be run using any Java-enabled browser at the URL http://www.ibc.wustl.edu/~ecr/CPG/segment.html. The purpose of the interface is to allow the user to input a nucleotide sequence in fasta format and then segment it into significant pieces based on the various compositions, the default of which is CpG islands. The results will be returned graphically to the user who can then analyze them interactively.

Two frames should initially appear when the applet is run. The first frame is the Sequence Fragmentation Interface frame (See figure 2) which is the main user interface. The second frame is the Status and Message Frame where error messages will be displayed as they occur. Other

messages will also appear in this frame in order to inform the user of the status of the breakpoint segmentation.

Setting the Parameters

Segment Composition. Clicking on an "Advanced Settings" menu, going to the "Segmentation Criteria" submenu, and clicking on the desired composition can change the criterion used for segmentation. There are currently five different compositions that can be used for segmentation criteria: mononucleotide, dinucleotide, CpG dinucleotide, purine/pyrimidine, and isochore (C+G) content.

Fasta Sequence File. The interface allows the user to input a DNA sequence in fasta format in one of three ways. One method is to input the sequence in a cut and paste fashion. A second method is to enter in a URL that points to a valid fasta file. The third method is to enter in the GenBank id number of a sequence.

Regardless of which method is chosen, a valid fasta file must be present. Fasta file format specifies that the first line begins with a '>' followed by the GSDB sequence accession number, the International Collaboration accession number, and a sequence description. The sequence follows the one line header. For the purposes of this segmentation program, it is only required that the first line begins with a '>'. Valid nucleotide characters of the sequence should follow the standard IUB/IUPAC nucleic acid codes as seen in table 2 (Moss, 1997). Note that the case of the characters can be mixed. In addition, spaces, tabs, and carriage returns are valid characters that will be stripped out prior to segmentation.

Note for the segmentation program, U will be converted to T, and anything besides A, C, G, or T will get set randomly according to the codes in table 2. If an invalid FASTA file is present, an error message will be displayed.

Minimum Threshold. The minimum threshold parameter allows the segmentation program to tell when segmentation should occur due to two segments being significantly different. If not enough breakpoints are appearing, lowering the threshold should introduce more. If too many breakpoints are appearing, then raise the threshold. A default value of 20 generally produces acceptable results. The user can change this value by changing the text box located to the right of the "Minimum Threshold" label. Note that this value is a real number.

Minimum Sequence Length. The minimum sequence length parameter refers to the minimum number of nucleotides that must be present in a segment. This parameter has been introduced, because without it, over segmentation becomes a problem. A default value of 100 is set. Updating the text box located to the right of the "Minimum Sequence Length" label can change this.

Post-processing. There is an additional post-processing parameter that can be set under the "Advanced Settings" menu. By checking the post-processing parameter, the segmentation program will attempt to merge breakpoints back together to form the most optimal results. This option is turned off by default.

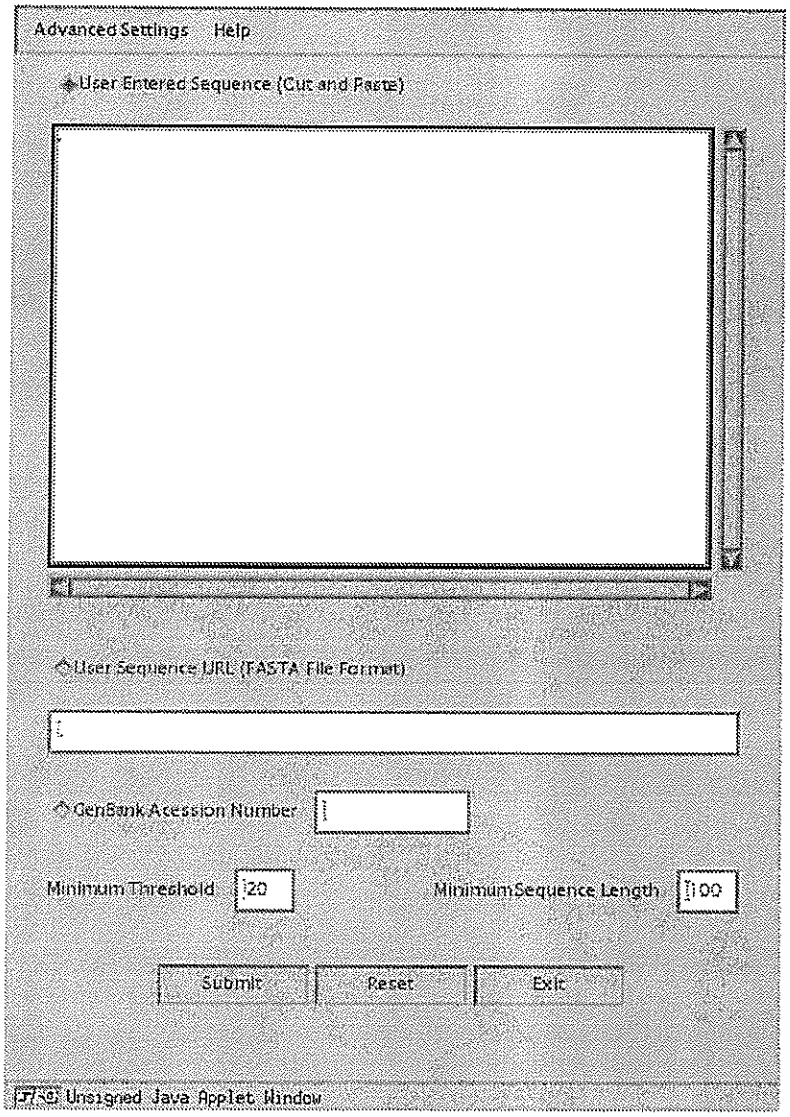

Figure 2: Sequence Fragmentation Interface

Symbol	Representation
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uridine
R	G A (purine)
Y	C T (pyrimidine)
K	G T (keto)

Symbol	Representation
M	A C (amino)
S	G C (strong)
W	A T (weak)
B	G T C
D	G A T
H	A C T
V	G C A
N	A G C T (any)

Table 2: IUB/IUPAC Nucleic Acid Codes

Interpretation of the Results

Once the breakpoint segmentation has occurred, two windows will pop up. One window indicates "Breakpoint Statistics" (figure 3) while the other is a "Choices" frame (figure 4.)

The window in figure 3 contains information on the nucleotide and possibly dinucleotide composition for each of the segments, as well as their beginning and ending points, their length, and the logP value. Included as well are the C + G composition, purine composition, and pyrimidine composition statistics for each of the segments. The title of the frame indicates the composition criteria for segmentation. In this case, segmentation is based on CpG content. When looking at the table of dinucleotide composition, the column labels refer to the first nucleotide and the row labels refer to the second nucleotide.

By clicking on any of the buttons in the Choices Frame as shown in figure 4, a graph will appear showing the content of the nucleotide(s) or dinucleotide(s) indicated on the button labels. Color-codes for the graphs are defined in table 3. Note that when multiple dinucleotides are shown together, the color corresponds to the second nucleotide.

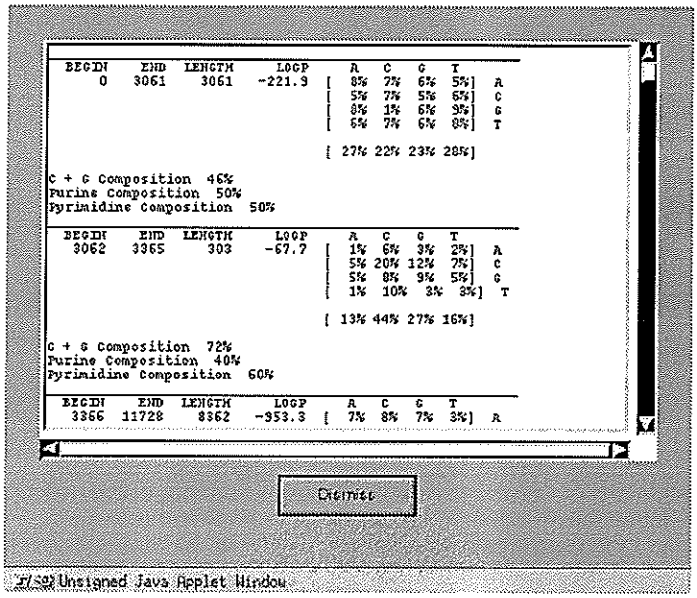

Figure 3: Breakpoint Statistics Frame

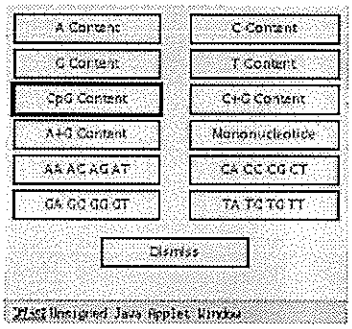

Figure 4: Choices Frame

Adenine	Green
Cytosine	Blue
Guanine	Black
Thymine	Red
All Others	Purple

Table 3: Nucleotide Color Codes

Figure 5 shows all of the breakpoints, which are indicated by the vertical dark blue lines. In this case, the breakpoints were determined according to CpG content. Note that the graphs are based on a running average over a specified window size. Editing the text to the right of the "Widow Size" label can change the window size. The graph will change according to the new window size once the "Redraw" button is pressed. The breakpoints might shift slightly to follow this window. If the graphs appear too cluttered, it would be best to increment the window size.

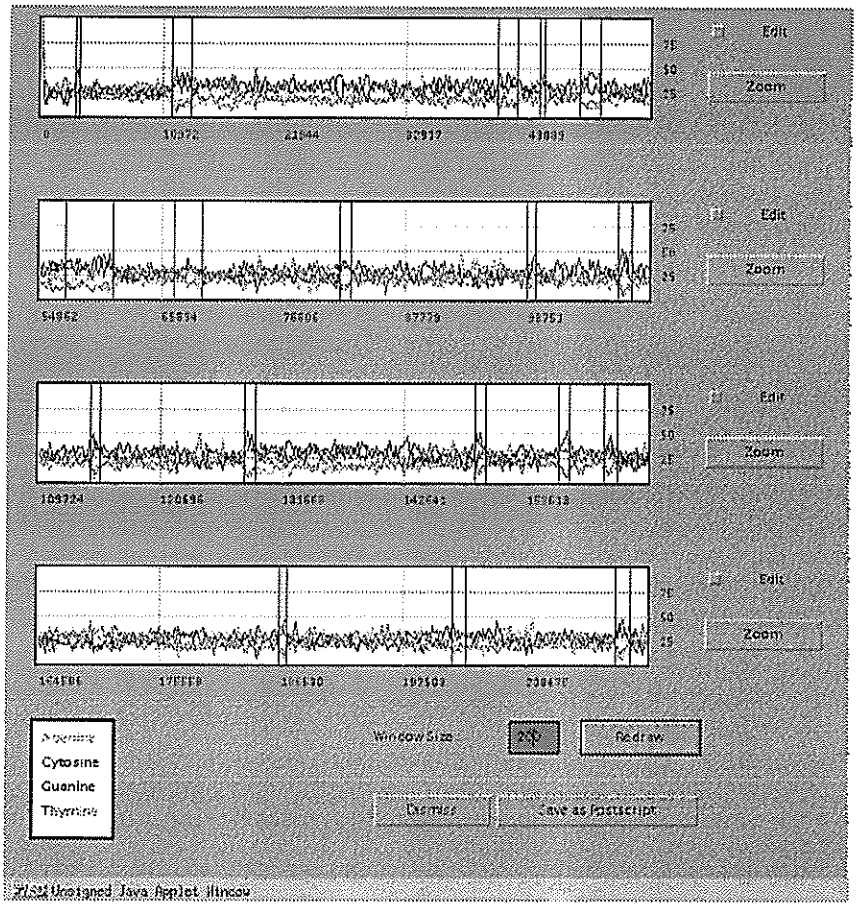

Figure 5: Mononucleotide Content using CpG Segmentation

Located directly to the right of each of the graphs is an "Edit" choice button. By clicking on this button, a blue background will appear on the associated graph. The user can then select a specific portion of the graph by either clicking or dragging the mouse to the desired location.

Once the desired area is covered, the user can press the associated zoom button to zoom in on this region of the graph.

Figure 6 shows an example of a zoomed in portion of a graph. The resulting zoom graph is very similar to the previous graphs. There are two main differences. The first is that when only a single composition is to be displayed, there will be blue tick marks underneath the graph indicating where it occurs within the sequence. The second difference is that there is a "View Sequence" button. By pressing this button, the nucleotide sequence will be displayed in a frame as shown in figure 7.

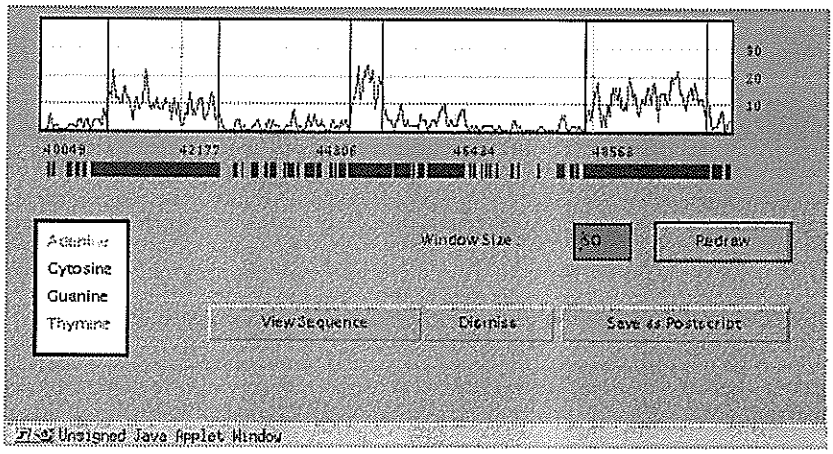

Figure 6: Zoom Graph of CpG Content

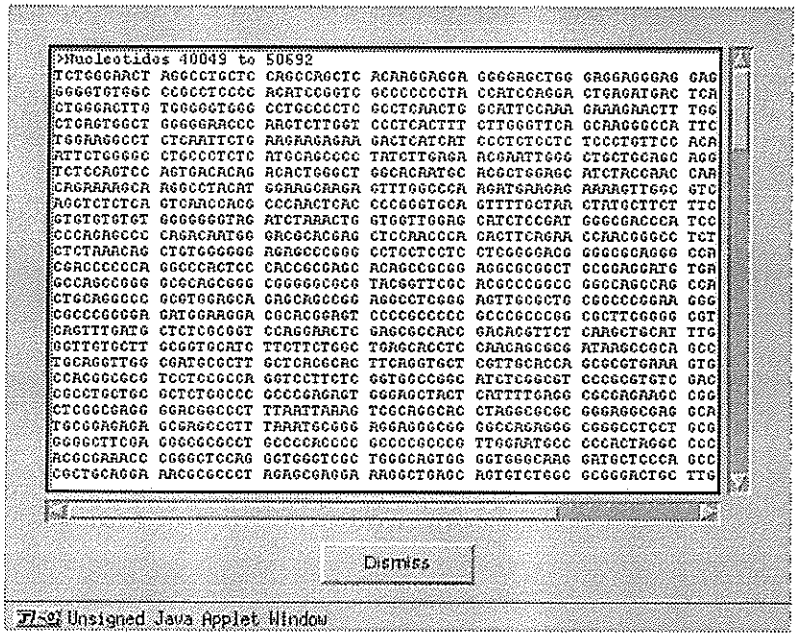

Figure 7: Nucleotide Sequence Frame

Implementation Issues

Due to the limitations of Java security, a client/server application is used in order to retrieve sequences from remote locations and to run the segmentation algorithm. Once the user has entered in the desired parameters in the client side applet, the parameters are sent to the server side executable. The server is responsible for taking in the parameters, retrieving the DNA sequence, and segmenting the sequence according to the provided parameters. Once the segmentation is finished, the server sends the results back to the client where they can be viewed graphically. All of the communication takes place through the use of sockets.

Code Statistics

For the client side Java applet, there are currently twenty-one different classes containing a total of 2104 lines of code. The server side consists of a single Java class that is 48 lines long and a C program that has 435 lines of code. The segmentation routines, written in C, take up six files containing a total of 943 lines of code.

Performing the actual segmentation in Java has been attempted, but is not feasible due to the nature of Java as an interpreted language. The bottleneck in the process is in I/O. Table 4 shows the runtime comparisons of the Java segmentation program versus the standalone executable created from compilation of C code. Testing was performed using a 55 MHz HyperSparc as the web server. The client side was run on a 200 MHz Pentium Pro machine. This data indicates that the Java interface slows down processing by a factor of 10.

Length (in Nucleotides)	JAVA Sequence Retrieval Time	JAVA Segmentation Time	Total JAVA Time	Standalone Segmentation Time
5828	17.3 Seconds	4.3 Seconds	21.6 Seconds	1.06 Seconds
93964	19.4 Seconds	9.3 Seconds	28.8 Seconds	2.33 Seconds
219446	36.0 Seconds	30.5 Seconds	66.5 Seconds	3.66 Seconds

Table 4 : Average Runtime Comparisons

Results

Figure 8 shows the breakpoint locations calculated within a sequence in the human Xq28 chromosomal region. The default parameters are used with the exception of the post-processing step being allowed. The location of breakpoints found is consistent with the results found by Chen, et. al. (1996). Our segmentation routine finds all of the CpG islands postulated, with an additional false positive island.

A subtle CpG island that cannot be picked out by the more traditional methods is shown in figure 9 for the bWXD3 region of the X chromosome. The minimum nucleotide length required for a segment is increased to 150. All other parameters take on their default values. Two CpG islands are postulated using these parameters. There is an exon located between bases 68432 and 68633 associated with the 3' end of the EDA gene. The first postulated CpG island is located between bases 85472 and 85727. This indicates that it is a good candidate located upstream of an exon associated with a gene. The second detected CpG island may indicate that there is another exon within this region.

Figure 10 shows the results for the bWXD42 region of the X chromosome that has a hint of a subtle CpG island. There is a cdx4 gene in this region with exons extending between bases 43025 (3' end) and 50304(5' end). Using the breakpoint segmentation program with a minimum threshold of 24 and default values for all of the other parameters, a single CpG island is located between bases 48716 and 50710. This indicates that that CpG island is actually located in the 5' end of the gene. More research will be pursued to determine the association between CpG island location and the 5' end of genes. The sequences for the bWXD3 and bWXD42 regions have been shared by the Washington University Medical School Center for Genetics in Medicine (CGM, 1997).

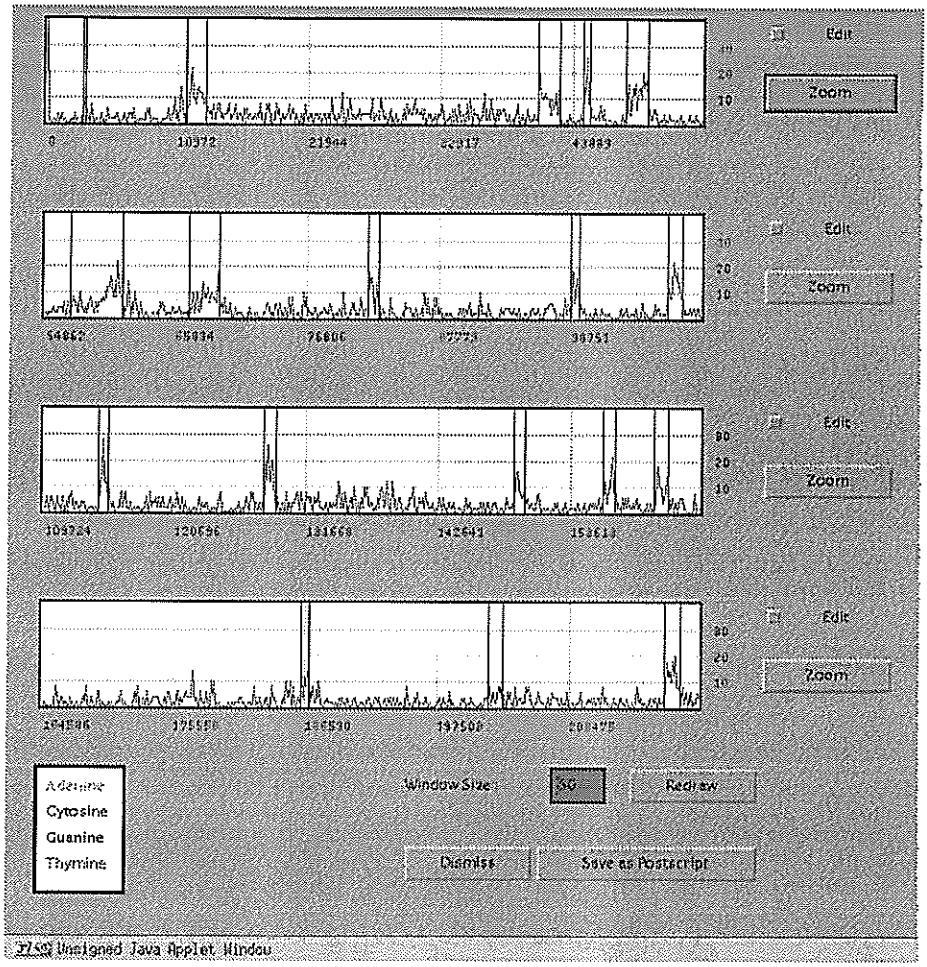

Figure 8: CpG Segmentation for Human Xq28 Chromosomal Region

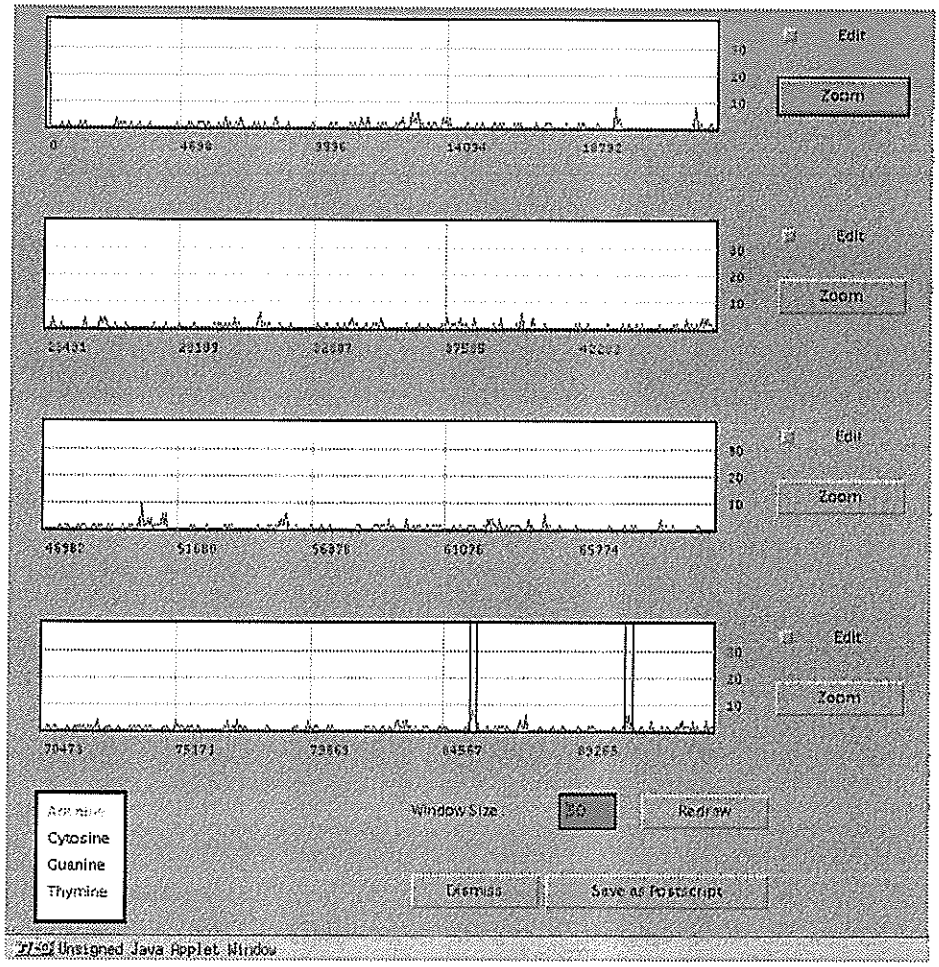

Figure 9: CpG Segmentation Results for bWXD3

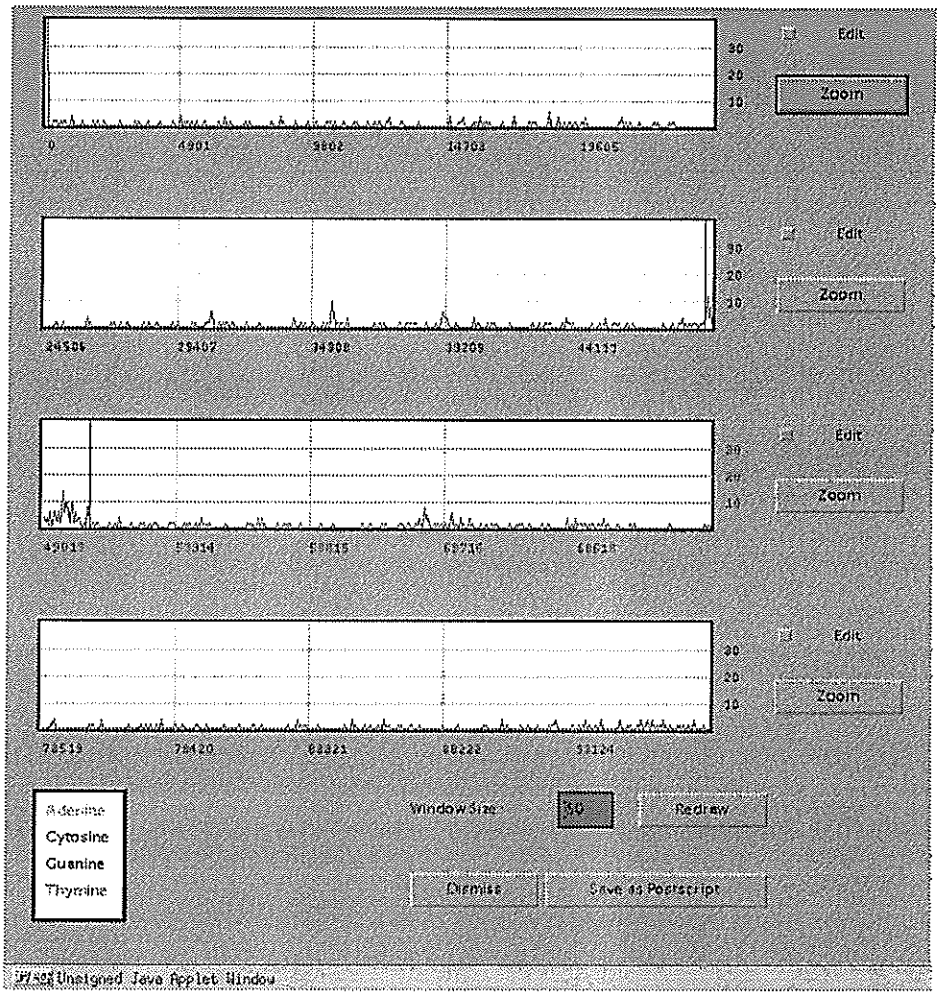

Figure 10: CpG Segmentation Results for bWXD42

Future Work

Currently, efforts have focused on the isolation of CpG islands for the purpose of analyzing the sequences to determine if there are any conserved sequence signals in either the beginning or end of the CpG islands that could lead to the conservation of CpG islands over the course of evolution. Gibbs sampling (Lawrence, et. al., 1993) and other similar motif identification programs will be used in this analysis. Other analysis will be performed to determine other conserved characteristics of CpG islands, including length, total CpG content, and location relative to the 5' start exons of genes.

There is room to improve the segmentation process. One area is to make a more accurate post-processing procedure to merge breakpoints without losing minor islands. Hopefully this would reduce false positives. We will also explore more analytical methods to determine segments taking segment length into account. Perhaps such a method will eliminate the need for a post-processing step.

The goal with the segmentation algorithm is to be able to develop an automatic method to annotate databases with added CpG island information. Hopefully this will add insight into the location of genes. While testing out the capabilities of this algorithm, it will be possible to assimilate a database of CpG islands more extensive than anything else currently available by looking at human DNA sequences extracted from GenBank and other databases.

Discussion of CpG islands has traditionally been limited to vertebrates. A comparison of homologous regions of DNA in mice and humans is possible. Through such a study, it can be determined which islands are conserved and which are lost. Future plans include analysis of other organisms including *S. cerevisiae* and *C. elegans* to determine if they have subtle CpG islands. Traditional methods suggest they do not.

The analysis performed so far suggests that there are at least 3 classifications of CpG islands: those having gradual signals, those having sharp left-handed signals, and those having sharp right-handed signals. A method using Kolmogorov-Smirnov testing (Lilliefors, 1967) will be explored in an attempt to classify CpG islands.

Segmentation can be applied to other sequence problems in addition to CpG island detection. The segmentation algorithm will be improved by allowing for the detection of other forms of compositional bias, introduction of higher-order oligomers, repeat sequences, and searching through amino acid sequences in addition to nucleic acid sequences.

Literature

Aissani, B. and Bernardi, G. CpG islands, genes and isochores in the genomes of vertebrates. *Gene*, **106**:185-195, 1991.

Antequera, F. and Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA*, **90**: 11995-19999, 1993.

Bernardi, G. The isochore organization of the human genome and its evolutionary history - a review. *Gene*, **135**:57-66, 1993.

BioTech Resources. "BioTech Life Science Dictionary" *Indiana Institute for Molecular and Cellular Biology*. 1996. <http://biotech.chem.indiana.edu/> (9 May 1997).

Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, **8**(7):1499-1504, 1980.

Bird, A.P., Functions for DNA Methylation in Vertebrates. *Cold Spring Harbor Symposia on Quantitative Biology*, **LVIII**:281-285, 1993.

Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, B.E., and Karlin, S. Methods and algorithms

for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA*, **89**:2002-2006, 1992.

Carlin, B.P., Gelfand, A.E., and Smith, A.F.M. Hierarchical Bayesian Analysis of Changepoint Problems. *Applied Statistics*, **41**(2):389-405, 1992.

CGM. "Center for Genetics in Medicine" *Washington University School of Medicine Center for Genetics in Medicine* . 1997. <http://genome.wustl.edu/cgm/> (9 May 1997).

Chen, C., Su, Y., Baybayan, P., Siruno, A., Nagaraja, R., Mazzarella, R. Schlessinger, D., Chen, E. Ordered shotgun sequencing of a 135 kb Xq25 YAC containing ANT2 and four possible genes, including three confirmed by EST matches. *Nucleic Acids Research*, **24**(20):4034-4041, 1996.

Chen, E.Y., Zolla, M., Mazzarella, R., Ciccodicola, A., Chen, C., Zuo, L., Heiner, C., Burough, F., Repetto, M., Schlessinger, D., D'Urso, M. Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Human Molecular Genetics*, **5**(5):659-668, 1996.

Cross, S.H. and Bird, A.P. CpG islands and genes. *Current Opinion in Genetics and Development*, **5**:309-314, 1995.

Curnow, R.N., and Kirkwood, T.B.L. Statistical Analysis of Deoxyribonucleic Acid Sequence Data – a Review. *Journal of the Royal Statistical Society A*, **152**:199-220, 1989.

Gardiner, K. Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends in Genetics*, **12**(12):519-524, 1996.

Larsen, F., Gundersen, G., and Lopez, L. CpG islands as Gene Markers in the Human Genome. *Genomics*, **13**:1095-1107, 1992.

Lawerence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science*, **262**: 208-214, 1993.

Lawrence, C.E. and Reilly, A.A. An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences. *PROTEINS: Structure, Function, and Genetics*, **7**:41-51, 1990.

Lawrence, C.E., and Reilly, A.A. Maximum Likelihood Estimation of Subsequence Conservation.
Journal of Theoretical Biology, **113**:425-439, 1985.

Lilliefors, H.W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. American Statistical Association Journal, **62**:399-402, 1967.

Lopez, R. "CpG Islands" *embnet.news 2(2) Database Development.* <http://biomaster.uio.no/cpgdb.html> (9 May 1997).

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes & Development*, **8**:2282-2292, 1994.

Moss, G.P. "Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences" *International Union of Pure and Applied Chemistry*. 1997. <http://alpha.qmw.ac.uk/~ugca000/iupac.html/misc/naseq.html> (2 Sep. 1997).

NCBI. "The National Center for Biotechnology Information" *The National Center for Biotechnology Information*. 1997. <http://www.ncbi.nlm.nih.gov/> (9 May 1997).

Rivella, S., Tamanini, F., Bione, S., Mancini, M. Herman, G., Chatterjee, A., Maestrini, E., and Toniolo, D. A Comparative Transcriptional Map of a Region of 250 kb on the Human and Mouse X Chromosome between the G6PD and the FLN1 Genes. *Genomics*, **28**:377-382, 1995.